

## SPAIN.

~	{	Llana.
		Rizada.
~	{	Marejadilla.
		Marejada.
~	{	Marejada gruesa.
		Gruesa.
~	{	Muy gruesa.
		Arbolada.
		Muy arbolada.

## D. CLOUD SYMBOLS.

Ley, W. Clement. *Cloudland*. London. 1894. p. 26-27.

Scientific name.	English name.
≡ Nebula.	Fog.
≡≡ Nebula pulverea.	Dust fog.
≡≡ Nebula stillans.	Wet fog.
= Nubes informis.	Scud.
= Stratus quietus.	Quiet cloud.
○ Stratus lenticularis.	Lenticular cloud.
≡ Stratus maculosus.	Mackerel cloud.
≡ Stratus castellatus.	Turret cloud.
≡ Stratus precipitans.	Plane shower.
○ Cumulo-rudimentum.	Rudiment.
○ Cumulus.	Heap cloud.
≡ Cumulo-stratus.	Anvil cloud.
≡ Cumulo-nimbus.	Shower cloud.
≡ Nimbus.	Rainfall cloud.
≡ Cumulo-stratus mammatus.	Tubercled anvil cloud.
≡ Cumulo-nimbus grandineus.	Hail shower.
≡ Cumulo-nimbus nivossus.	Snow shower.
≡ Cumulo-nimbus mammatus.	Festooned shower cloud.
≡ Nimbus grandineus.	Hail-fall.
≡ Nimbus nivossus.	Snow-fall.
∩ Nubes fulgens.	Luminous cloud.
∩ Cirrus.	Curl cloud.
∩ Cirro-filum.	Gossamer cloud.
∩ Cirro-velum.	Veil cloud.
∩ Cirro-macula.	Speckle cloud.
∩ Cirro-velum mammatum.	Draped veil cloud.

Howard, Luke. *On the modifications of clouds*. London. 1803. p. 14. (Hellmann's "Neudrucke," No. 3, Berlin, 1894.)

- ∩ Cirrus.
- Cumulus.
- Stratus.
- ∩ Cirro-cumulus.
- ∩ Cirro-stratus.
- ∩ Cumulo-stratus.
- ∩ Cirro-cumulo-stratus, or Nimbus.

Formerly used by Iowa Weather Service. (Adopted 1876.)

- ∩ Cirrus.
- ∩ Cirro-stratus.
- ∩ Cirro-cumulus.
- ∩ Cumulus.
- ∩ Pallio-cirrus.
- ∩ Pallio-cumulus.
- ∩ Fracto-cumulus.
- = Polar bands, drawn as placed across the sky with → indicating motion; thus ∩→ bands NW-SE moving toward the east.

## E. LITERAL SYMBOLS.

In addition to arbitrary symbols, numerous literal symbols—usually the initial letter or letters of meteorological terms in various languages—have been used in meteorological registers and on weather maps. Only a few of these are included in the foregoing lists. The rest lie beyond the scope of the present compilation.

## ON THE COEFFICIENT OF CORRELATION AS A MEASURE OF RELATIONSHIP.

By CHARLES N. MOORE.

[Dated: University of Cincinnati, Department of Mathematics, Apr. 17, 1916.]

In recent years several applications of the theory of correlation have been made in connection with meteorological investigations.<sup>1</sup> Consequently a brief discussion of the significance of a correlation coefficient and its reliability as a measure of relationship may be of interest to readers of the MONTHLY WEATHER REVIEW. The theoretical discussion in the present paper is in substance the same as that given by the writer in a recent paper in *Science*.<sup>2</sup> The bearing of that discussion on applications in meteorology is given here for the first time.

The theory of correlation deals with the relationship between two variable quantities whose variations are due in part or entirely to common causes. A certain quantity,  $r$ , known as a coefficient of correlation, is computed, and from its value inferences are drawn as to the extent to which the variations of the two quantities are affected in the same way by the same causes, or as to the extent to which the variation of one quantity affects that of the other.

The formula for  $r$  in terms of  $n$  pairs of observed values of two variables  $x$  and  $y$ , is

$$r = \frac{\sum_{i=1}^{i=n} (x_i - x_0) (y_i - y_0)}{\sqrt{\sum_{i=1}^{i=n} (x_i - x_0)^2 \cdot \sum_{i=1}^{i=n} (y_i - y_0)^2}}, \quad (1)$$

where  $x_0$  is the mean of the  $x$  values and  $y_0$  the mean of the  $y$  values.<sup>3</sup> The value of  $r$  obtained from this formula

<sup>1</sup> See J. Warren Smith in MONTHLY WEATHER REVIEW, February, 1914, 42:78; and *ibid.*, 1915, 43:222.

A. Sresnevsky, in *Meteorologische Zeitschrift*, Braunschweig, December, 1914, 31: 506.

L. Steiner, in *Meteorologische Zeitschrift*, Braunschweig, September, 1915, 32: 419.

<sup>2</sup> Moore, *Chas. N.* On the coefficient of correlation as a measure of relationship. *Science*, New York, October 22, 1915 (NS), 42:575-579.

<sup>3</sup> For an account of the process of computing  $r$  from a table of observed values of two variables see the paper by J. Warren Smith, MONTHLY WEATHER REVIEW, February, 1914, 42:79-80.

will never be less than  $-1$  nor greater than  $+1$ , and in general will have a value lying between these two values. In case the variations of the two quantities depend entirely on common causes in such a way that one variable can be expressed in terms of the other by means of an equation,  $r$  may take on one of the extreme values  $+1$  or  $-1$ . It will take on one of these values if one variable can be expressed linearly in terms of the other; i. e., if

$$y = ax + b,$$

where  $a$  and  $b$  are constants. In this case it will be  $+1$  if  $a > 0$  and  $-1$  if  $a < 0$ ; in all other cases it will have a value lying between these two values. In case the two variables are entirely independent of each other in the sense that their variations have no common causes,  $r$  will be zero or very near to zero if the number of observed values of  $x$  and  $y$  is large enough to eliminate the effects of chance. If the value of  $r$  lies between  $0$  and  $-1$  or between  $0$  and  $+1$ , it may be due to the fact that there is a relation between the two variables that is not linear, or to the fact that the variations of the two quantities are not due to common causes in such a way that one can be expressed in terms of the other alone.

In general the variable quantities under discussion will have their variations subject to a great variety of causes. Let us assume, then, that

$$\begin{aligned} x &= f_1(\epsilon_1, \epsilon_2, \dots, \epsilon_m), \\ y &= f_2(\epsilon_1, \epsilon_2, \dots, \epsilon_m), \end{aligned}$$

where  $\epsilon_1, \epsilon_2, \dots, \epsilon_m$  are  $m$  independent variables, and  $f_1$  and  $f_2$  are two different expressions in terms of those variables. If we are to be able to give any sort of definite interpretation to the value of  $r$ , it is necessary to assume further that the  $f$ 's are, to a good degree of approximation, linear expressions in the  $\epsilon$ 's, i. e., that the equations

$$\begin{aligned} x &= a_{11}\epsilon_1 + a_{12}\epsilon_2 + \dots + a_{1m}\epsilon_m, \\ y &= a_{21}\epsilon_1 + a_{22}\epsilon_2 + \dots + a_{2m}\epsilon_m, \end{aligned} \quad (2)$$

where the  $a$ 's are constants, are approximately true. If now we represent the deviation of each  $\epsilon$  from its mean value by a  $v$  with the corresponding subscript, we obtain readily from the last equations

$$\begin{aligned} x - x_0 &= a_{11}v_1 + a_{12}v_2 + \dots + a_{1m}v_m, \\ y - y_0 &= a_{21}v_1 + a_{22}v_2 + \dots + a_{2m}v_m, \end{aligned} \quad (3)$$

where  $x_0$  and  $y_0$  are the mean values of  $x$  and  $y$ , respectively. It is evident that the mean value of each  $v$  is zero, since each represents the deviation of the corresponding  $\epsilon$  from its mean value.

Suppose now that  $(v_i', v_j')$ ,  $(v_i'', v_j'')$ ,  $\dots$ ,  $(v_i^{(n)}, v_j^{(n)})$  are  $n$  pairs of observed values of  $v_i$  and  $v_j$ . Since the  $\epsilon$ 's are independent variables, we shall have if  $n$  is very large,

$$\sum_{r=1}^{r=n} v_i^{(r)} v_j^{(r)} = 0. \quad (4)$$

For in that case there will be associated with each particular value of  $v_i$  a series of values of  $v_j$  whose mean is zero. Hence, if we collect terms involving the same values of  $v_i$  the sum of each set of terms will be zero, and therefore the whole summation in (4) will reduce to zero. For values of  $n$  small enough to make the computation of  $r$  practicable, equation (4) will in general be only approximately true. The larger  $n$  is, the better in general the approximation will be.

We will now substitute the values of  $x - x_0$  and  $y - y_0$  given by (3) in (1) and take account of (4) in making the substitution. We obtain

$$r = \frac{\sum_{i=1}^{i=n} a_{1i} a_{2i} s_i^2}{\sqrt{\sum_{i=1}^{i=n} a_{1i}^2 s_i^2} \sqrt{\sum_{i=1}^{i=n} a_{2i}^2 s_i^2}}, \quad (5)$$

where we have set

$$s_i^2 = \frac{v_i'^2 + v_i''^2 + \dots + v_i^{(n)2}}{n} \quad (i = 1, 2, \dots, m). \quad (6)$$

The  $s$ 's thus defined are known as the *standard deviations* of the corresponding  $\epsilon$ 's. The formula (5) for  $r$  is well adapted to the discussion of the connection between the value of  $r$  and the degree of relationship between  $x$  and  $y$ .

To illustrate the way in which we can interpret the significance of  $r$  by means of equation (5) we will consider a particular example. Suppose the two variables  $x$  and  $y$  represent the wind velocities measured at the same instant in two different localities.<sup>4</sup> If the localities are not too far apart it is reasonable to suppose that the two velocities will depend in part on the same causes. Such a state of affairs will be represented mathematically by the equations in (2) if we suppose that the  $a$ 's in the first equation corresponding to a certain set of the  $\epsilon$ 's are zero, and the  $a$ 's in the second equation corresponding to a different set of  $\epsilon$ 's are zero.

We will suppose then that the first  $p$  of the  $a$ 's in the first equation are zero and the last  $q$  of those in the second equation are zero, i. e., that

$$\begin{aligned} a_{11} &= a_{12} = \dots = a_{1p} = 0, \\ a_{2, m-q+1} &= a_{2, m-q+2} = \dots = a_{2m} = 0. \end{aligned} \quad (7)$$

In order to begin with a fairly simple example we will suppose further that we are dealing with a case where each of the other  $a$ 's involved is equal to a single positive quantity  $a$ , i. e., where

$$\begin{aligned} a_{1, p+1} &= a_{1, p+2} = \dots = a_{1m}, \\ a_{21} &= a_{22} = \dots = a_{2, m-q} = a > 0. \end{aligned} \quad (8)$$

It is readily seen that the  $s$ 's defined by equation (6) depend upon the scales used in the measurement of the different  $\epsilon$ 's. Therefore there will be no loss of generality in supposing that these scales are so chosen that each of the  $s$ 's is equal to a single quantity  $s$ , i. e., that

$$s_1 = s_2 = s_3 = \dots = s_m = s. \quad (9)$$

If we substitute the values given by (7), (8), and (9) in (5), we obtain

$$r = \frac{(m-p-q)a^2s^2}{\sqrt{(m-p)a^2s^2} \sqrt{(m-q)a^2s^2}} = \frac{m-p-q}{\sqrt{(m-p)(m-q)}}. \quad (10)$$

We shall now make use of (10) to indicate one way in which the value of  $r$  throws light upon the relationship

<sup>4</sup> The correlation between two such variables is considered in the paper by Sreanewsky referred to above.

between a variation in  $x$  and a corresponding variation in  $y$ . Let us suppose that all the  $\epsilon$ 's on which  $x$  depends, i. e., all the  $\epsilon$ 's for which the corresponding  $a$ 's on the right-hand side of the first equation in (1) are not zero, are increased by a certain quantity  $d$ , whereas all the other  $\epsilon$ 's, i. e., all the  $\epsilon$ 's on which  $y$  depends but  $x$  does not depend remain constant. If we represent by  $x'$  the new value of  $x$ , and by  $y'$  the new value of  $y$  after the increase in the  $\epsilon$ 's, then when we take account of equations (7), (8), and (2) we have readily

$$\begin{aligned} y' - y &= (m - p - q)ad, \\ x' - x &= (m - p)ad. \end{aligned} \quad (11)$$

Since the units in terms of which  $x$  and  $y$  are measured are in general arbitrary, it is apparent that we need to introduce some standard unit for each of them before we can attach any definite significance to a comparison of their changes in value. The natural way to choose a unit for this purpose is to relate its size in some definite way to the range of variability of the variable quantity concerned. This can be done by choosing as a unit the *standard deviation* of each variable. The standard deviations of the  $\epsilon$ 's, as stated above, are given by equation (6). The standard deviation of any other variable is defined in an analogous manner. Hence in view of equations (3), (4), (6), (7), (8), and (9), we have for the standard deviations of  $x$  and  $y$

$$s_x = \sqrt{\frac{m-p}{n}} as, \quad s_y = \sqrt{\frac{m-q}{n}} as. \quad (12)$$

$R$  may be said to be a good measure of the closeness of relationship between the two variables since it measures the extent to which a *typical* change in one variable causes a corresponding change in the other variable. If now we set

$$R = \frac{(y' - y)/s_y}{(x' - x)/s_x}, \quad (13)$$

we obtain from (11) and (12)

$$R = \frac{m - p - q}{\sqrt{(m - p)(m - q)}}.$$

Hence in this particular case  $r = R$ , and  $r$  may therefore be said to be a good measure of the degree of relationship between  $x$  and  $y$ .

It is easy to see, however, that cases may arise in which  $r$  and  $R$  differ considerably in value. Suppose, for example, that the  $a$ 's of equation (2) satisfy the following conditions:

$$\begin{aligned} a_{11} &= a_{12} = \dots a_{1p} = a_{2,m-p+1} = \dots a_{2m} = 0, \\ a_{21} &= a_{22} = \dots a_{2p} = a_{1,m-p+1} = \dots a_{1m} = 10a, \\ a_{i,p+1} &= \dots a_{i,p+2} = \dots a_{i,m-p} = a. \quad (i = 1, 2, \dots) \\ (m &= 102p) \end{aligned}$$

Then we find by substituting in formulæ (5) and (13) that

$$r = 0.5, \quad R = 0.9.$$

Under still other suppositions the discrepancy between the values of  $r$  and  $R$  may be still greater. Hence it is apparent that  $r$  may not always be a good measure of the closeness of relationship between two variable quantities.

The chief conclusion to be drawn from the foregoing discussion is to a considerable extent a negative one. It is shown that it is possible to state conditions under which the coefficient of correlation as calculated from equation (1) will furnish a reliable measure of the degree of relationship between two variable quantities. But it is also shown that in cases where these conditions are not approximately fulfilled, the coefficient of correlation will not necessarily be a good measure of this relationship. As there seems to be no way of determining in any particular case whether or not the conditions we have stated are satisfied, it is apparent that considerable caution should be observed in drawing definite inferences from the value of a coefficient of correlation.

5-13 (5/13) 1916/11

## RAINFALL IN CHINA, 1900-1911.<sup>1</sup>

By CO-CHING CHU, A. M.

[Dated: Cambridge, Mass., Mar. 21, 1916.]

### INTRODUCTION.

As the fluctuation in rainfall from year to year is great, it is always a difficult matter to discuss the subject and draw isohyets with accuracy and intelligence unless we have a long series of reliable observations well distributed over the region under discussion.

China has been backward on all subjects meteorological. The data on rainfall in China are mostly spasmodic, inaccurate, and limited to recent years only. The data on rainfall in this article are based on Rev. Louis Froc's work "La Pluie en Chine, durant une période de onze années, 1900-1911," published by the Catholic Mission of Zi-ka-wei, Shanghai, China. These are, no doubt, the most recent and at the same time the most reliable data on the rainfall in China. In all, there are 88 stations, divided into four classes according to the length of the record of rainfall. In the first class, which comprises 34 stations, all except 4 have data extending through the period of 11 years. The records of the remaining stations are incomplete, varying in length from eight to two or three years. The stations are not very well distributed, but are concentrated mostly along the coast and the valley of the Yangtze River; in the northwest they are entirely wanting. The area of China proper, according to Mill's International Geography, is approximately 1,300,000 square miles. Assuming that all the data of the 88 stations were available and that they were uniformly distributed, there still would be only one station to every 1,500 square miles.

It is evident that a rainfall map based upon these data can only be tentative. If the stations were more numerous and better distributed, and if the records extended over a longer period, the map would be probably quite different from what it is.

### RAINFALL CONTROLS.

In the main, there are three factors which control the amount and seasonal distribution of precipitation in China, (1) the monsoon, (2) the topography, and (3) the cyclonic distribution.

(1) *Monsoon*.—The monsoon<sup>2</sup> is a seasonal wind which is best developed in Asia, owing to the vastness of

<sup>1</sup> A study offered as part of the requirements for the degree of A. M. at Harvard University in 1915; prepared under the direction of Prof. A. G. McAdie and R. De C. Ward.  
<sup>2</sup> Whether the summer southeast wind in China should be called "monsoon" or "trade wind" is controversial according to Mr. B. C. Wallis. See the extract from a paper by him, MONTHLY WEATHER REVIEW, January, 1915, 43:24.